




# Using national electronic health records for pandemic preparedness: validation of a parsimonious model for predicting excess deaths among those with COVID-19—a data-driven retrospective cohort study

Mehrdad A Mizani<sup>1,2</sup>, Ashkan Dashtban<sup>1</sup>, Laura Pasea<sup>1</sup>, Alvina G Lai<sup>1</sup>, Johan Thygesen<sup>1</sup>, Chris Tomlinson<sup>1</sup> , Alex Handy<sup>1</sup>, Jil B Mamza<sup>3</sup>, Tamsin Morris<sup>3</sup>, Sara Khalid<sup>4</sup>, Francesco Zaccardi<sup>5</sup>, Mary Joan Macleod<sup>6</sup>, Fatemeh Torabi<sup>7</sup>, Dexter Canoy<sup>8</sup>, Ashley Akbari<sup>7</sup> , Colin Berry<sup>9</sup>, Thomas Bolton<sup>2</sup>, John Nolan<sup>2</sup>, Kamlesh Khunti<sup>5</sup>, Spiros Denaxas<sup>1</sup>, Harry Hemingway<sup>1</sup>, Cathie Sudlow<sup>2</sup> and Amitava Banerjee<sup>1</sup> ;

on behalf of the CVD-COVID-UK Consortium

<sup>1</sup>Institute of Health Informatics, University College London, London NW1 2DA, UK

<sup>2</sup>BHF Data Science Centre, Health Data Research UK, London, NW1 2BE, UK

<sup>3</sup>Medical and Scientific Affairs, BioPharmaceuticals Medical, AstraZeneca, Cambridge, CB2 0AA, UK

<sup>4</sup>Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, OX3 7HE, UK

<sup>5</sup>Leicester Diabetes Centre, University of Leicester, Leicester, LE5 4PW, UK

<sup>6</sup>School of Medicine, Medical Sciences and Nutrition, University of Aberdeen, Aberdeen, AB24 3FX, UK

<sup>7</sup>Faculty of Medicine, Health and Life Science, Swansea University, Swansea, SA2 8QA, UK

<sup>8</sup>Nuffield Department of Women's and Reproductive Health, University of Oxford, Oxford, OX3 9DU, UK

<sup>9</sup>Institute of Cardiovascular and Medical Sciences, University of Glasgow, Glasgow, G12 8TA, UK

**Corresponding author:** Amitava Banerjee. Email: [ami.banerjee@ucl.ac.uk](mailto:ami.banerjee@ucl.ac.uk)

## Summary

**Objectives:** To use national, pre- and post-pandemic electronic health records (EHR) to develop and validate a scenario-based model incorporating baseline mortality risk, infection rate (IR) and relative risk (RR) of death for prediction of excess deaths.

**Design:** An EHR-based, retrospective cohort study.

**Setting:** Linked EHR in Clinical Practice Research Datalink (CPRD); and linked EHR and COVID-19 data in England provided in NHS Digital Trusted Research Environment (TRE).

**Participants:** In the development (CPRD) and validation (TRE) cohorts, we included 3.8 million and 35.1 million individuals aged  $\geq 30$  years, respectively.

**Main outcome measures:** One-year all-cause excess deaths related to COVID-19 from March 2020 to March 2021.

**Results:** From 1 March 2020 to 1 March 2021, there were 127,020 observed excess deaths. Observed RR was 4.34% (95% CI, 4.31–4.38) and IR was 6.27% (95% CI, 6.26–6.28). In the validation cohort, predicted one-year excess deaths were 100,338 compared with the observed 127,020 deaths with a ratio of predicted to observed excess deaths of 0.79.

**Conclusions:** We show that a simple, parsimonious model incorporating baseline mortality risk, one-year IR and RR of the pandemic can be used for scenario-based prediction of excess deaths in the early stages of a pandemic. Our

analyses show that EHR could inform pandemic planning and surveillance, despite limited use in emergency preparedness to date. Although infection dynamics are important in the prediction of mortality, future models should take greater account of underlying conditions.

## Keywords

Clinical, epidemiology, health informatics, infectious diseases, public health

Received: 16th June 2022; accepted: 24th September 2022

## Introduction

Mortality estimates of COVID-19 have been widely reported and followed at local, regional, national and international levels since early in the pandemic, influencing policy and health service planning. Electronic health record (EHR) data informed early identification of risk factors for COVID-19 severity and mortality, leading to UK lockdown and shielding policies.<sup>1–3</sup> Moreover, EHR linkage enabled both specialist registry data and pragmatic clinical trials of new treatments at scale.<sup>4,5</sup>

All-cause and disease-specific mortality prediction in research and clinical practice has included underlying conditions or ‘baseline mortality risk’, often derived and validated using EHR.<sup>6–8</sup> Underlying non-communicable diseases (NCDs) are important mortality predictors in infectious diseases,<sup>9,10</sup> but baseline mortality risk based on NCDs is largely neglected in pandemic preparedness, which emphasises infection transmissibility and severity, using metrics such as case fatality ratio, infection fatality ratio and reproduction number.<sup>11–14</sup> Although COVID-19 is increasingly viewed as a ‘syndemic’<sup>15</sup> (with interaction between infectious diseases and NCDs, requiring cross-speciality expertise), efforts to predict excess mortality have focused on dynamic transmission modelling without consideration of baseline risk or use of anonymised, individual-level, population-scale EHR.<sup>16,17</sup>

On 22 March 2020, before the first UK lockdown, we released a preprint (published on 12 May 2020),<sup>1</sup> estimating one-year COVID-19 mortality using a model developed in pre-pandemic population-based linked EHRs from 3.8 million people in the UK (via Clinical Practice Research Datalink [CPRD]). Our EHR-derived model included baseline one-year mortality risk for a range of underlying conditions, incorporating scenario-based assumptions regarding relative risk (RR) of mortality during the pandemic compared to baseline, and population infection rate (IR). Validation of the model is required to establish the actual RR and IR, to update scenario-based assumptions and to assess the accuracy of model predictions.

The NHS Digital Trusted Research Environment (TRE) for England, which became available during 2020, offers the opportunity to validate our approach at whole population level, with longitudinal, individual-level data.<sup>18,19</sup> Therefore, using these data, we: (1) ascertained the observed IR of COVID-19 and RR of one-year COVID-19 mortality; and (2) compared the predicted versus observed COVID-19 mortality for conceptual validation of our EHR-derived model.

## Methods

### Data sources

**Conceptual model development.** We used a pre-pandemic linked CPRD dataset, including EHR, across primary care, hospital data and death registry with follow-up from 1997 to 2017.<sup>1</sup>

**Model validation.** The NHS Digital TRE for England provides secure, remote access to linked, individual-

level EHR data,<sup>18,19</sup> including primary care, hospital episodes, registered deaths, dispensed medicines, COVID-19 laboratory tests and vaccinations. We used General Practice Extraction Service Data for Pandemic Planning and Research, Hospital Episode Statistics Admitted Patient Care, Second Generation Surveillance System, COVID-19 Hospitalisation in England Surveillance System, Civil Registry Deaths, NHS Business Services Authority dispensed medicines and COVID-19 vaccine datasets, prior to 15 May 2021.<sup>19</sup>

### Cohort specifications

Both model development and validation involved population-based, retrospective cohort analyses with a range of high-risk conditions as exposures and one-year all-cause mortality as outcome. In the validation study, a further exposure was SARS-CoV-2 infection. In the development study, eligible individuals were aged  $\geq 30$  years, registered with a GP between 1 January 1997 and 1 January 2017 (Figure S1.A), with  $\geq 1$  year of follow-up.

In the validation study, eligible individuals were aged  $\geq 30$  years on 1 March 2018. COVID-19-related high-risk conditions were from Public Health England guidance.<sup>20</sup> We considered all-cause mortality after COVID-19 as the direct pandemic effect. Deaths in those without COVID-19 include baseline mortality and deaths attributable to indirect pandemic effects. To evaluate direct COVID-19 effects on one-year all-cause mortality, we specified two time periods (Figure S1.B and S1.C). The pre-pandemic period (1 March 2018–1 March 2019) was used for baseline characteristics and outcome (mortality) in the non-exposed (non-COVID-19) group. The pandemic period (1 March 2020–1 March 2021) was used to study COVID-19 cases and deaths in the exposed group (i.e. COVID-19 with or without high-risk conditions). Underlying conditions were assessed on 1 March 2018 in the validation study, minimising the effect of age difference between pre-pandemic and pandemic periods (Figure S2).

### Exposures and outcomes of interest

Exposures were presence (versus absence) of high-risk conditions for COVID-19<sup>20</sup> including cardiovascular disease (CVD), chronic kidney disease (CKD), diabetes, chronic obstructive pulmonary disease (COPD), body mass index (BMI) over 40 kg/m<sup>2</sup>, chronic liver disease, age  $>70$  years and history of oral steroid therapy. For all conditions, except steroid therapy, the minimum period between earliest diagnosis date and baseline date (1 March 2018)

was one year. For steroid therapy, event date was based on first dispensing date between 1 March 2018 and 1 March 2019, since prescription/dispensed medication data were only available since April 2018. Outcome was one-year all-cause mortality.

To define underlying conditions, we used extended CALIBER phenotyping algorithms.<sup>21</sup> Phenotypes with earliest diagnosis dates between 1 March 2017 and 1 March 2018 were excluded, to allow  $\geq 1$  year history of conditions prior to cohort entry. The CVD phenotype was a composite, including heart failure, stroke (non-specified, ischaemic, haemorrhagic, transient ischaemic attack, subarachnoid haemorrhagic), arrhythmias, acute myocardial infarction, cardiomyopathy, atrial fibrillation, deep vein thrombosis, isolated calf vein thrombosis and pulmonary embolism. The dispensed oral corticosteroid phenotype was determined based on the CALIBER phenotype mapped to British National Formulary codes.<sup>22</sup> To define COVID-19 cases, we used positive swab testing results and Public Health England labs and NHS hospitals, community swab testing results, primary care and hospital episode data, vaccination and death registration.<sup>23</sup>

### Model development and validation

Our prediction model in the development study was a conceptual model based on baseline mortality, RR of death in those exposed to COVID-19 versus those not exposed to COVID-19 (pre-pandemic) and IR of COVID-19

$$\frac{\text{COVID-19-related all-cause excess death count}}{\text{Baseline death count}} = \text{IR}(\text{RR} - 1)$$

In the development study, we calculated scenario-based COVID-19 excess deaths using baseline mortality by high-risk underlying conditions and plausible RR/IR (0.001%, 1%, 10% and 80% for total, partial, moderate and no suppression, respectively).<sup>2</sup> For each IR scenario, we applied RRs (1.2, 1.5 and 3), and scaled up to mid-2018 population of England aged  $\geq 30$  using estimates of the Office for National Statistics.<sup>24</sup>

Full validation was beyond scope and not possible in the rapidly changing timelines of the pandemic. Validation in our study involved use of observed IR and RR values (TRE for England; Figure S1.B) in the conceptual model to predict COVID-19 deaths in the development and validation cohorts. This constituted ‘model verification’ (‘determining that the model’s inputs and outputs are consistent with actual data and accepted theories’) and ‘conceptual

model validation’ (‘determining that the theories and assumptions underlying the conceptual model are correct and the model representation of the problem entity and the model’s structure, logic and mathematical and causal relationships are reasonable for the intended purpose of the model’).<sup>25</sup> In order to capture direct COVID-19 mortality effects, we selected unexposed and exposed groups in pre-pandemic and pandemic periods, respectively. We estimated baseline one-year mortality in the pre-pandemic period (Figure S1.B) by Kaplan–Meier survival analysis. We calculated baseline and COVID-19 mortality risk (using RR) in pre-pandemic and pandemic periods, respectively, by high-risk conditions. To calculate IR in each sub-sample, we divided the COVID-19 population by those at risk at the start of the period. The final IR was the average of IRs of two sub-samples (refer to Supplementary materials).

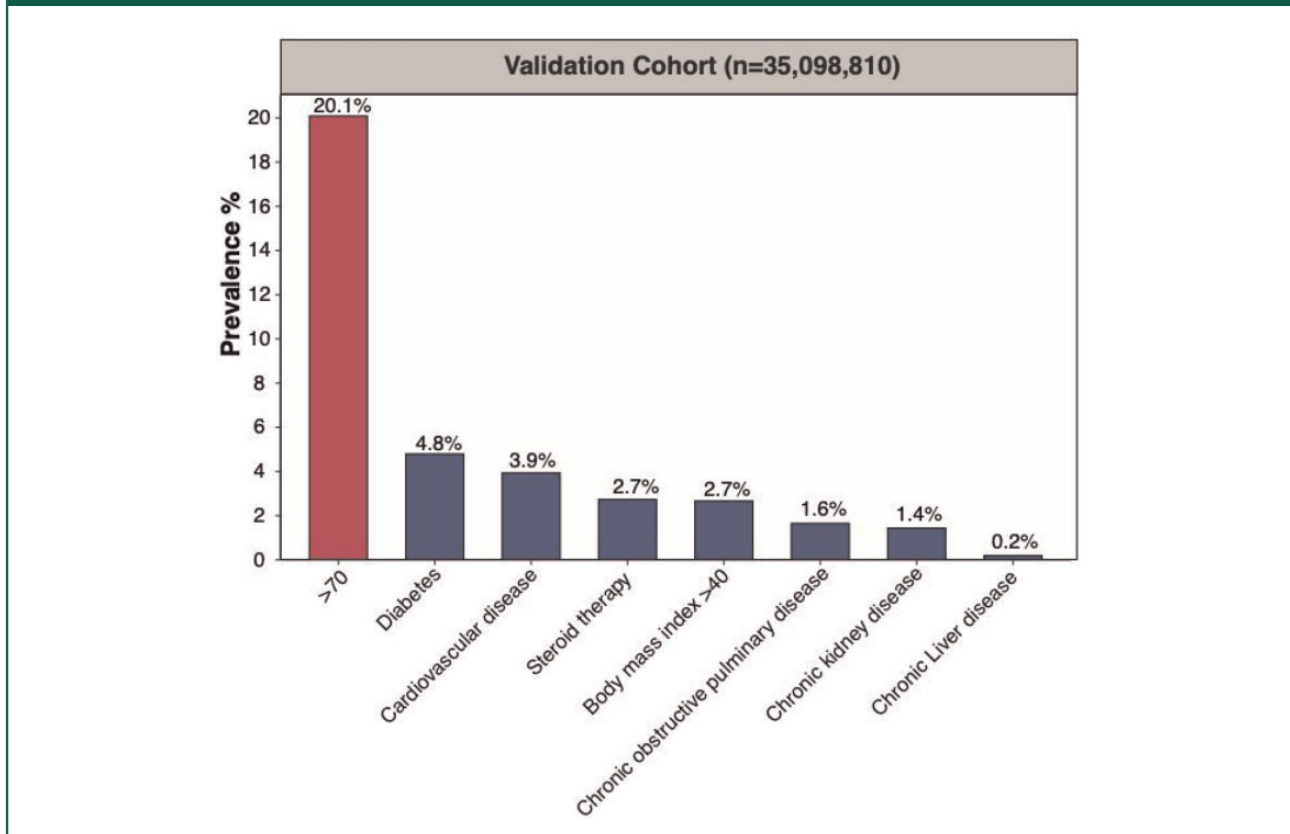
### Results

In the validation cohort, we included 35,098,810 individuals aged  $\geq 30$  years at baseline (Figure S2). Of all individuals aged  $\geq 30$  years on 1 March 2018, 18,361,665 (52.3%) were female; mean age was 55.0 [SD 16.2] in both sexes; 28,049,984 (79.9%) were aged  $\leq 70$  (mean age 48.7 [SD 11.6] years in females and 49.1 [SD 11.5] years in males) and 7,048,826 (20.1%) were  $> 70$  (mean 79.7 [SD 6.8] years in females and 78.5 [SD 6.1] years in males). Prevalence for CVD, diabetes, CKD, COPD, BMI  $> 40$  kg/m<sup>2</sup>, chronic liver disease and steroid therapy was 5.56% and 2.76%, 4.59% and 3.75%, 2.03% and 2.84%, 1.83% and 1.81%, 1.41% and 2.07%, 0.15% and 0.10%, and 3.52% and 5.07% in males and females, respectively. Prevalence of 0, 1, 2 and  $\geq 3$  underlying conditions was 35.57% and 39.95%, 8.15% and 8.48%, 8.82% and 2.79%, and 1.13% and 1.09% in males and females, respectively. Prevalence of all underlying conditions was higher in individuals  $> 70$  years and males (Figure 1, Table 1).

### One-year mortality

Among individuals with at least one high-risk condition, estimated pre-pandemic one-year mortality risk was observed to be 3.55% (3.54–3.57). One-year mortality risk in individuals aged  $> 70$  years was 9.24% (9.17–9.31), 3.37% (3.34–3.40), 8.36% (8.32–8.40) and 6.38% (6.34–6.42) for COPD, CKD, CVD and diabetes, respectively. In individuals aged  $> 70$  years, one-year mortality risks in men were 9.45% (9.35–9.55), 3.91% (3.85–3.96), 7.92% (7.98–9.20), 6.48% (6.42–6.54) for COPD, CKD, CVD and diabetes, respectively; and in women, 9.02% (8.92–9.11),

**Figure 1.** Prevalence of high-risk conditions for COVID-19 mortality in validation cohort (n = 35,098,810) cohort aged  $\geq 30$  years.



3.00% (2.96–3.04), 8.84% (8.78–9.11) and 6.27% (6.21–6.33), respectively.

### Validation and replication of the conceptual model

In March 2020, we predicted 73,498 one-year COVID-19-related deaths for the population of England, by scaling from the development cohort (3,862,012 aged  $\geq 30$  years) to the mid-2018 population of England and assuming a scenario of  $IR = 10\%$  and  $RR = 3$ .<sup>2</sup> In the validation study, from March 2020 until March 2021, we ascertained 127,020 COVID-19-related all-cause deaths. We estimated pre-pandemic one-year mortality risk by age group, sex and number of high-risk conditions in the absence of COVID-19.

We calculated cross-validated one-year (March 2020–2021)  $RR$  and  $IR$  of COVID-19 as 4.34% (95% CI, 4.31–4.38) and 6.27% (95% CI, 6.26–6.28), respectively. Tables S1 and S2 show the cross-validated  $IR$  and  $RR$ , respectively, across two random subsamples of the cohort shown in Figure S1. Table S3 shows the sensitivity analysis for under-fitting and further cross-validation. We found that

the effect of vaccination on overall  $RR$  or  $IR$  between December 2020 and March 2021 was negligible compared to effects of under-reported COVID-19 cases pre-vaccination (Table S4). We applied our prediction model using observed  $RR$  (4.34) and  $IR$  (6.27) and baseline mortality risk data in the validation cohort (Tables S5 and S6).

Figures 3 and S4 show the predicted one-year COVID-19-related all-cause deaths, based on baseline mortality risk (March 2018–2019 for validation cohort),  $RR = 4.34$ , and  $IR = 6.27\%$  compared to observed excess deaths (March 2020–2021). The observed and model-predicted COVID-19 deaths were 127,020 and 100,338 (79.0% of observed), respectively (Tables 2, Figure 3).

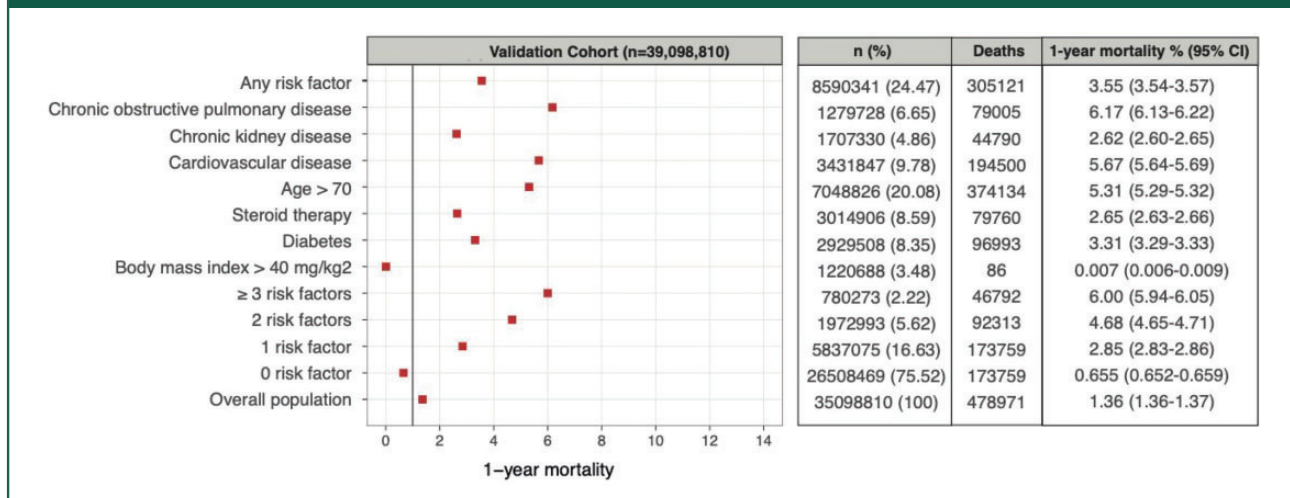
### Discussion

In anonymised, individual-level, population-scale, national EHR data between March 2020 and March 2021, we conducted the first study to predict and validate one-year mortality among those with COVID-19 using baseline (pre-pandemic) mortality risk. We provide the first detailed, scenario-based

**Table 1.** Underlying conditions in the validation cohort (NHS Digital TRE, n = 35,098,810, aged 30 years or older).

Underlying condition	Count (% of total population)					
	Male		Female		All ages	
	Age ≤ 70 years N = 13,587,089	Age > 70 years N = 3,150,056	Age ≤ 70 years N = 14,462,895	Age > 70 years N = 3,898,770	Female Age > 70 years N = 3,898,770	Female All ages N = 18,361,665
CVD	873,001 (2.49)	1,080,487 (3.08)	1,953,488 (5.56)	509,450 (1.45)	968,909 (2.76)	1,478,359 (4.21)
Diabetes	965,436 (2.75)	647,269 (1.84)	1,612,705 (4.59)	716,309 (2.04)	600,494 (1.71)	1,316,803 (3.75)
CKD	227,924 (0.65)	483,972 (1.38)	711,896 (2.03)	274,852 (0.78)	720,582 (2.05)	995,434 (2.84)
COPD	291,294 (0.83)	351,684 (1.00)	642,978 (1.83)	287,287 (0.82)	349,463 (0.99)	636,750 (1.81)
BMI > 40 kg/m <sup>2</sup>	373,213 (1.06)	120,512 (0.34)	493,725 (1.41)	561,351 (1.60)	165,612 (0.47)	726,963 (2.07)
Chronic liver disease	42,789 (0.12)	10,966 (0.03)	53,755 (0.15)	25,807 (0.07)	9875 (0.03)	35,682 (0.10)
Steroid therapy	762,449 (2.17)	472,571 (1.35)	1,235,020 (3.52)	1,183,308 (3.37)	596,578 (1.70)	1,779,886 (5.07)
0	11,167,965 (31.82)	1,317,372 (3.75)	12,485,337 (35.57)	12,137,332 (35.58)	1,885,800 (5.37)	14,023,132 (39.95)
1	1,835,747 (5.23)	1,025,674 (2.92)	2,861,421 (8.15)	1,800,831 (5.13)	1,174,823 (3.35)	2,975,654 (8.48)
2	451,492 (1.29)	541,211 (1.54)	992,703 (2.82)	406,504 (1.16)	573,786 (1.63)	980,290 (2.79)
≥3	131,885 (0.38)	265,799 (0.76)	397,684 (1.13)	118,228 (0.34)	264,361 (0.75)	382,589 (1.09)

**Figure 2.** Baseline one-year mortality in England (age  $\geq 30$  years) according to underlying conditions in the validation cohort (n = 35,098,810).



mortality risk assessment before and during the pandemic, based on absolute risk estimates in national population data. We show that a simple, parsimonious model incorporating baseline risk of mortality, IR and RR of the pandemic can be used to predict one-year COVID-19 mortality.

### Strengths and weaknesses

Our analysis uses anonymised, national, individual-level EHR data with unprecedented scale and whole population inclusivity and validated EHR phenotypes. It highlights the importance of EHR data, baseline mortality and scenario-based assumptions in risk assessment at early stages of a pandemic where dynamics of the new infectious disease are not yet known.

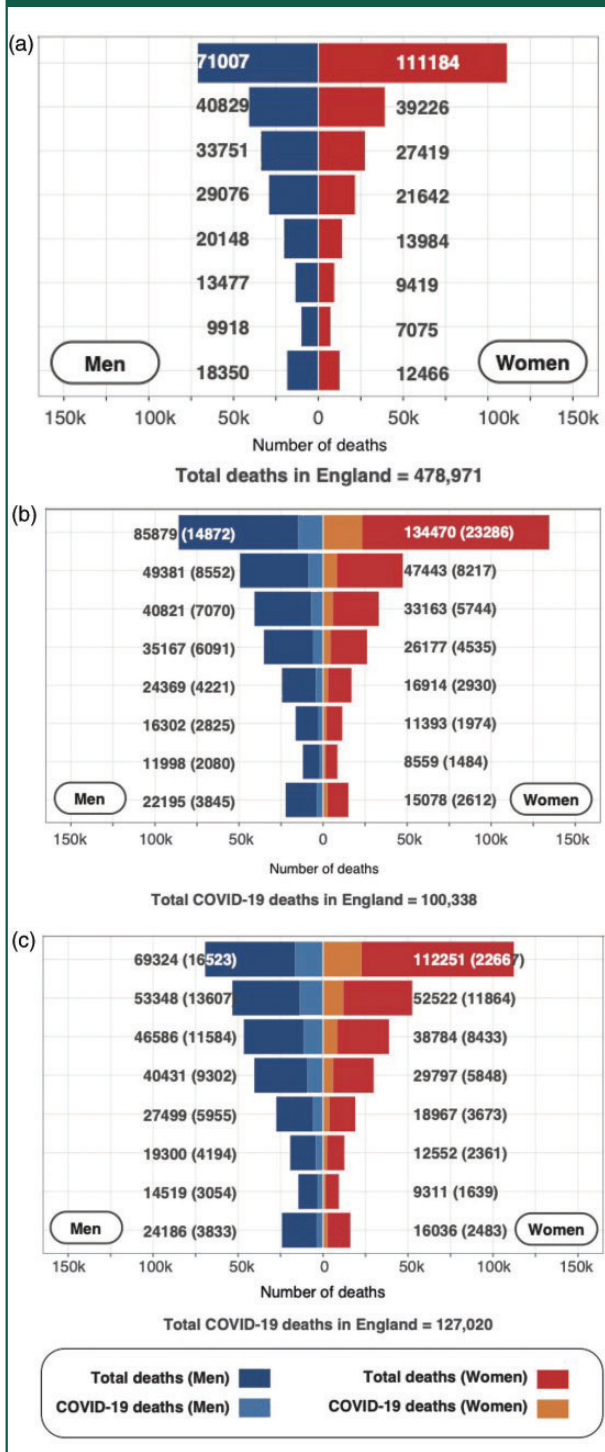
Our analysis used only the most frequent high-risk conditions. Our simple model made assumptions regarding static RR and IR over the course of the pandemic and did not incorporate infectivity or population dynamics of the original or later strains of SARS-CoV-2, the impact of COVID-19-related policies or vaccination rates. Generalisability of our findings to other countries and contexts requires further validation. Our study only investigated COVID-19, and applicability to other infectious diseases or pandemics is unknown. There are differences between development and validation cohorts in terms of data coding systems (e.g. lack of standardised one-to-one mapping between coding terminologies), and limited availability of fields in CPRD (e.g. ethnicity) and in the TRE for England (e.g. medication use before 2018 and multiple index of deprivation), which restricted analyses. Overall, national

mortality estimates in people with COVID-19 were similar in development and validation cohorts, with differences in mortality risk at baseline in stratified analyses. For example, mortality risk was similar for younger people in both cohorts, but mortality risk was relatively higher in the development cohort for individuals aged >70 years due to the earlier cohort entry date in the CPRD study population.<sup>1</sup> Also, the number of estimated deaths was lower in the development cohort in all age categories, perhaps because the one-year mortality in CPRD data were calculated after study entry date, when these individuals were younger (mean age 43.5 [SD 11.7] years), compared to the validation cohort in March 2018–2019 (mean age 55.0 [SD 16.2] years). Another explanation is that the actual IR over one year is higher than our observed rate (and probably greater than the 10% we used in prediction), due to incomplete availability of COVID-19 testing, especially during the early months of the pandemic.

### Comparison with other studies

We searched for systematic reviews published after March 2020 in PubMed using combinations of equivalent Mesh terms: ‘COVID’, ‘prediction’, ‘mortality’, ‘model’, ‘underlying condition’, ‘relative risk’ and ‘infection rate’. A systematic review of 107 multivariate prediction models for COVID-19 mortality showed that variables were selected from signs, symptoms and risk factors from COVID-19 patients during the pandemic.<sup>26</sup> All models had unclear or high risk of bias, including non-representative data sources, unreliable COVID-19 case definition, excluding patients who had not experienced

**Figure 3.** Baseline deaths, model-predicted COVID-19-related all-cause deaths and observed deaths among those with COVID-19 in England (age  $\geq 30$  years) over one year, stratified by age and sex in the validation cohort (n = 35,098,810). (a) Baseline one-year mortality. (b) Total (and model-predicted COVID-19) one-year mortality based on RR = 4.34%, IR = 6.27%. (c) Total (and observed COVID-19) one-year mortality.



outcomes of interest, and model overfitting. We found no studies of excess mortality prediction based on pre-pandemic mortality in people with high-risk underlying conditions and RR and IR associated with COVID-19. In our study, all patients, regardless of outcome of interest, were included in analyses. Moreover, we conducted model cross-validation to minimise overfitting (Table S3).

We used EHR data of the whole population in England to validate our model for predicting one-year excess mortality in people exposed to COVID-19. The data used in our study are derived from anonymised, individual-level and linked EHRs of the whole population in England, making our model highly representative. We have used validated phenotype definitions for high-risk underlying conditions and COVID-19 cases. Our study highlights the significance of pre-pandemic longitudinal EHR data to predict the direct effects of the pandemic for preparedness and early response.

Our model is a simple, conceptual model for formulating worst-case to best-case scenarios at the start of the pandemic. We developed the model in CPRD data with assumed parameters and replicated the model in NHS Digital TRE using observed RR and IR values. Hence, our model is more suitable for risk assessment for pandemic preparedness and early response rather than high-precision estimation of the mortality.

### Meaning of the study: possible mechanisms and implications

**Pre-pandemic mortality risks.** Baseline mortality risk can be used to predict COVID-19-related mortality over one year at the national level, and underlying conditions and age are major determining factors of the risk. We show that national data EHRs, such as the NHS Digital TRE, and sampled less complete data, such as CPRD, can be used to estimate and monitor baseline risk at scale. Such data are available across diseases, risk factors and countries via the Global Burden of Disease Study and other efforts, and have already been used to project high-risk populations for COVID-19.<sup>27</sup> There is public demand for such information, which can be provided in an interpretable, usable format employing open phenotypes, coding and standards.<sup>18–21,28</sup>

**Infection rate over one year.** Surveillance of SARS-CoV-2 IRs has been crucial across countries throughout the pandemic by different methods, including incident or prevalent cases, over weeks or months, by antigen or antibody tests, or by static or dynamic rates. Our model used population IR over one year,

**Table 2.** Observed COVID-19 one-year mortality in England (NHS Digital TRE; n = 35,098,810 aged ≥30 years; 1 March 2020 to 1 March 2021).

	Age ≤70 years			Age >70 years			All ages					
	N total (%)	Total deaths	N COVID-19 (%)	COVID-19 deaths	N total (%)	Total deaths	N COVID-19 (%)	COVID-19 deaths	N total (%)	Total deaths	N COVID-19 (%)	COVID-19 deaths
≥1 Underlying condition excluding age >70 years	4,634,608 (13.58)	70,202	314,587 (0.92)	16,203	3,340,209 (9.79)	317,114	209,190 (0.61)	74,276	7,974,817 (23.36)	70,479	523,777 (1.53)	90,479
Age >70 years	–	–	–	–	6,299,844 (18.46)	443,043	317,798 (0.93)	99,828	–	–	–	–
Diabetes	1,645,037 (4.82)	29,688	123,984 (0.36)	8338	1,087,148 (3.18)	106,124	75,870 (0.22)	27,474	2,732,158 (8.00)	135,902	199,854 (0.58)	35,812
CVD	1,335,614 (3.91)	30,301	80,174 (0.23)	6966	1,710,348 (5.01)	200,644	121,772 (0.36)	46,744	3,045,962 (8.92)	230,945	201,946 (0.59)	53,710
BMI > 40 kg/m <sup>2</sup>	932,120 (2.73)	8454	73,399 (0.21)	2333	280,331 (0.82)	19,410	15,690 (0.04)	4911	1,212,451 (3.55)	27,864	89,089 (0.26)	7244
Steroid therapy	1,889,695 (5.54)	44,671	149,685 (0.44)	7655	923,584 (2.70)	111,144	67,321 (0.20)	24,354	2,813,279 (8.24)	155,815	217,006 (0.64)	32,009
COPD	549,304 (1.61)	18,905	29,797 (0.09)	3733	574,369 (1.68)	70,701	43,183 (0.13)	16,872	112,3673 (3.29)	89,606	72,980 (0.21)	20,605
CKD	492,763 (1.44)	11,102	33,377 (0.10)	3255	1,100,918 (3.25)	121,830	75,622 (0.22)	29,332	1,593,680 (4.67)	132,932	108,999 (0.32)	32,587
Chronic liver disease	60,270(0.18)	3584	3769 (0.18)	556	15,556 (0.04)	2291	1213 (0.003)	483	75,826(0.22)	5875	4982(0.01)	1039
3+ Underlying conditions	233,799 (0.68)	12,645	18,267 (0.05)	1470	442,569 (1.30)	67,507	40,625 (0.12)	17,304	676,368 (1.98)	80,152	58,892 (0.17)	20,774
2 Underlying conditions	827,803 (2.472)	20,516	55,977 (0.16)	4885	956,907 (2.80)	104,452	66,645 (0.19)	24,693	1,784,710 (5.23)	12,4968	122,622 (0.36)	29,578
1 underlying condition	3,573,006 (10.47)	37,041	240,343 (0.70)	7848	1,940,733 (5.68)	145,155	101,920 (0.30)	32,279	5,513,739 (16.15)	182,196	342,263 (1.00)	40,127
No underlying condition	23,197,624 (47.96)	72,168	1,615,026 (4.73)	10,989	2,959,635 (8.67)	125,929	108,608 (0.32)	23,869	26,157,259 (76.63)	198,097	1,723,634 (5.05)	36,541
Overall population	27,832,232 (81.54)	142,370	1,929,613 (5.65)	27,192	6,299,844 (18.46)	443,043	317,798 (0.93)	99,828	34,132,076 (100)	585,413	2,247,411 (6.58)	127,020



which we estimated using comprehensive testing, primary care, hospital data and death data in the NHS Digital TRE in a mostly pre-vaccination era. Our estimates of IR represent nearly the whole English population, consistent with pre-vaccination antibody rates in the UK<sup>29</sup> and a recent study using the same data.<sup>23</sup> However, under-estimation is still possible and, moreover, likely, due to initially limited testing capacity and asymptomatic infection. Future research and models should incorporate higher vaccination rates, novel variants, potential impact of reinfection and dynamic IRs over time.

**Relative risk associated with the pandemic.** Excess mortality associated with COVID-19 has been a focus in health policy since the early stages of the pandemic. Comparisons with flu persist until now, including ‘winter excess deaths’, which have been estimated as 20% higher than the baseline mortality rate.<sup>1</sup> In our model, we used RR estimates of 1.5, 2 and 3, and in the national data, we observed 4.34 in the overall population. Assuming an under-estimation of IR, we may have over-estimated RR, but our estimates are in line with a recent time-series analysis of excess mortality in the first pandemic wave in the UK. That study showed that certain underlying conditions were associated with higher RR of excess pandemic mortality, compared with the pre-pandemic period.<sup>30</sup>

### **Implications for public health and policy makers**

There are three public health and policy implications. First, EHRs were designed and used for reimbursement, clinical care and quality improvement, with limited use in emergency preparedness. Our analyses show that EHRs could and should be part of pandemic planning and surveillance. Second, pre-pandemic mortality risk can be estimated at individual, subgroup and national levels, and is important in pandemic mortality prediction as well as preparedness including shielding and vaccination prioritisation. Third, our data support the syndemic lens, which views COVID-19 not just as an infectious disease, but one with social, environmental and NCD determinants and effects, signalling the need for multidisciplinary public health and policy approaches in pandemics.

**Research implications.** First, there are more than 80 diseases, risk factors and underlying conditions designated as moderate and high risk for COVID-19 by the UK government.<sup>20</sup> We will validate COVID-19 mortality estimates for the comprehensive list, providing condition-specific IR and RR estimates, stratified by ethnicity, deprivation and vaccination, with

future application for models in COVID-19 and other pandemics. Second, the policy need for region- and country-specific data is well recognised, and our UK-based analyses may not be generalisable to other countries and datasets. Third, we only considered direct pandemic impact on mortality, not indirect and long-term (Long COVID) impact, which need to be studied and incorporated into future pandemic impact models. Fourth, baseline mortality risk estimation (using models such as ours) could be combined with existing methods of dynamic transmission modelling to predict and mitigate future pandemics.

### **Conclusions**

The impact of the COVID-19 pandemic on excess mortality can be predicted using national EHRs and is related to baseline mortality risk, population IRs and pandemic-associated RR. In public health, policy and research, there are implications for expertise, data and resources in future pandemic preparedness.

### **Declarations**

**Competing Interests:** The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: JBM and TM are employees of AstraZeneca. KK is chair of the ethnicity subgroup of the Independent Scientific Advisory Group for Emergencies (SAGE) and director of the University of Leicester Centre for Black Minority Ethnic Health. KK and AB are trustees of the South Asian Health Foundation (SAHF). CS is Director of the BHF Data Science Centre. All other authors report no competing interests.

**Funding:** The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The British Heart Foundation Data Science Centre (grant no. SP/19/3/34678, awarded to Health Data Research (HDR) UK) funded co-development (with NHS Digital) of the TRE, provision of linked datasets, data access, user software licences, computational usage, and data management and wrangling support, with additional contributions from the HDR UK data and connectivity component of the UK Government Chief Scientific Adviser’s National Core Studies programme to coordinate national COVID-19 priority research. Consortium partner organisations funded the time of contributing data analysts, biostatisticians, epidemiologists and clinicians. AB, MAM, MHD and LP were supported by research funding from AstraZeneca. AB has received funding from the National Institute for Health Research (NIHR), British Medical Association and UK Research and Innovation. AB, SD and HH are part of the BigData@Heart Consortium, funded by the Innovative Medicines Initiative-2 Joint Undertaking under grant agreement No 116074. KK is supported by the National Institute for Health Research (NIHR) Applied Research Collaboration East Midlands (ARC-EM) and NIHR Lifestyle BRC.

**Ethics approval:** Approval for the study in CPRD was granted by the Independent Scientific Advisory Committee (20\_074R) of the Medicines and Healthcare products Regulatory Agency in the

UK in accordance with the Declaration of Helsinki. The North East-Newcastle and North Tyneside 2 research ethics committee provided ethical approval for the CVD-COVID-UK research programme (REC No 20/NE/0161).

**Information governance:** The data used in this study are available in NHS Digital's TRE for England, but as restrictions apply, they are not publicly available (<https://digital.nhs.uk/coronavirus/coronavirus-data-services-updates/trusted-research-environment-service-for-england>). The CVD-COVID-UK/COVID-IMPACT programme led by the BHF Data Science Centre (<https://www.hdrk.ac.uk/helping-with-health-data/bhf-data-science-centre/>) received approval to access data in NHS Digital's TRE for England from the Independent Group Advising on the Release of Data (<https://digital.nhs.uk/about-nhs-digital/corporate-information-and-documents/independent-group-advising-on-the-release-of-data>) via an application made in the Data Access Request Service Online system (ref. DARS-NIC-381078-Y9C5K) (<https://digital.nhs.uk/services/data-access-request-service-dars/dars-products-and-services>). The CVD-COVID-UK/COVID-IMPACT Approvals & Oversight Board (<https://www.hdrk.ac.uk/projects/cvd-covid-uk-project/>) subsequently granted approval to this project to access the data within NHS Digital's TRE for England. The de-identified data used in this study were made available to accredited researchers only.

The open-source code and utilised phenotype code-lists used in this study are available in a repository in the British Heart Foundation Data Science Centre's GitHub organisation ([https://github.com/BHFDSC/CCU003\\_03](https://github.com/BHFDSC/CCU003_03)).

**Guarantor:** AB.


**Contributorship:** Research question, approach and study oversight: AB. Leading data engineering, coding and analysis: MAM. Data analysis, quality assurance and phenotyping: AD, JT, CT, AH, TB, JN. Study design and review: LP, SD, HH, CS, MJM, DC, CB, KK. Data visualisation: AGL. Coordinating approval for and access to data within NHS Digital's TRE for England for CVD-COVID-UK/COVID-IMPACT: CS. Drafting initial and final versions of manuscript: AB and MAM. Critical review of early and final versions of manuscript: All authors.

**Acknowledgements:** This work is carried out with the support of the BHF Data Science Centre led by HDR UK (BHF Grant no. SP/19/3/34678) and makes use of de-identified data held in NHS Digital's TRE for England, made available via the BHF Data Science Centre's CVD-COVID-UK/COVID-IMPACT consortium. This work uses data provided by patients and collected by the NHS as part of their care and support. We would also like to acknowledge all data providers who make health relevant data available for research.

**Provenance:** Not commissioned; peer reviewed by Julie Morris, and pre-submission peer review comments.

**ORCID iDs:** Chris Tomlinson  <https://orcid.org/0000-0002-0903-5395>

Ashley Akbari  <https://orcid.org/0000-0003-0814-0801>

Amitava Banerjee  <https://orcid.org/0000-0001-8741-3411>

**Supplemental material:** Supplemental material for this article is available online.

## References

1. Banerjee A, Pasea L, Harris S, Gonzalez-Izquierdo A, Torralbo A, Shallcross L, et al. Estimating excess

1-year mortality associated with the COVID-19 pandemic according to underlying conditions and age: a population-based cohort study. *Lancet* 2020; 395: 1715–1725.

2. Williamson EJ, Walker AJ, Bhaskaran K, Bacon S, Bates C, Morton CE, et al. Factors associated with COVID-19-related death using OpenSAFELY. *Nature* 2020; 584: 430–436.
3. Clift AK, Coupland CAC, Keogh RH, Diaz-Ordaz K, Williamson E, Harrison EM, et al. Living risk prediction algorithm (QCOVID) for risk of hospital admission and mortality from coronavirus 19 in adults: national derivation and validation cohort study. *BMJ* 2020; 371: m3731.
4. Docherty AB, Harrison EM, Green CA, Hardwick HE, Pius R, Norman L, et al. Features of 20 133 UK patients in hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: prospective observational cohort study. *BMJ* 2020; 369: m1985.
5. Horby P, Lim WS, Emberson JR, Mafham M, Bell JL, Linsell L, et al.; RECOVERY Collaborative Group. Dexamethasone in hospitalized patients with Covid-19. *N Engl J Med* 2021; 384: 693–704.
6. Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, Minhas R, Sheikh A, et al. Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. *BMJ* 2008; 336: 1475–1482.
7. Vogelsang RP, Bojesen RD, Hoelmich ER, Orhan A, Buzquurz F, Cai L, et al. Prediction of 90-day mortality after surgery for colorectal cancer using standardized nationwide quality-assurance data. *BJS Open* 2021; 5: zrab023.
8. Ajnakina O, Agbedjro D, McCammon R, Faul J, Murray RM, Stahl, et al. Development and validation of prediction model to estimate 10-year risk of all-cause mortality using modern statistical learning methods: a large population-based cohort study and external validation. *BMC Med Res Methodol* 2021; 21: 1–11.
9. Bolge SC, Kariburyo F, Yuce H and Fleischhackl R. Predictors and outcomes of hospitalization for influenza: real-world evidence from the United States medicare population. *Infect Dis Ther* 2021; 10: 213–228.
10. Ma HM, Tang WH and Woo J. Predictors of in-hospital mortality of older patients admitted for community-acquired pneumonia. *Age Ageing* 2011; 40: 736–741.
11. Huppert A and Katriel G. Mathematical modelling and prediction in infectious disease epidemiology. *Clin Microbiol Infect* 2013; 19: 999–1005.
12. Biggerstaff M, Cowling BJ, Cucunubá ZM, Dinh L, Ferguson NM, Gao H, et al. WHO COVID-19 modelling parameters group. Early insights from statistical and mathematical modeling of key epidemiologic parameters of COVID-19. *Emerg Infect Dis* 2020; 26: e1–e14.
13. Laydon DJ, Mishra S, Hinsley WR, Samartsidis P, Flaxman S, Gandy A, et al. Modelling the impact of the tier system on SARS-CoV-2 transmission in the

- UK between the first and second national lockdowns. *BMJ Open* 2021; 11: e050346.
14. Verity R, Okell LC, Dorigatti I, Winskill P, Whittaker C, Imai N, et al. Estimates of the severity of coronavirus disease 2019: a model-based analysis. *Lancet Infect Dis* 2020; 20: 669–677.
  15. Horton R. Offline: COVID-19 is not a pandemic. *Lancet* 2020; 396: 874.
  16. Banerjee A, Chen S, Pasea L, Lai AG, Katsoulis M, Denaxas S, et al. Excess deaths in people with cardiovascular diseases during the COVID-19 pandemic. *Eur J Prev Cardiol* 2021; 28: 1599–609.
  17. Lai AG, Pasea L, Banerjee A, Hall G, Denaxas S, Chang WH, et al. Estimating excess mortality in people with cancer and multimorbidity in the COVID-19 emergency. *BMJ Open* 2020; 10: e043828.
  18. Wood A, Denholm R, Hollings S, Cooper J, Ip S, Walker V, et al. Linked electronic health records for research on a nationwide cohort of more than 54 million people in England: data resource. *BMJ* 2021; 373: n826.
  19. CVD-COVID-UK/COVID-IMPACT TRE Dataset Provisioning Dashboard, British Heart Foundation Data Science Centre, Health Data Research UK. See [www.hdruk.ac.uk/wp-content/uploads/2022/02/220210-CVD-COVID-UK-COVID-IMPACT-TRE-Dataset-Provisioning-Dashboard.pdf](http://www.hdruk.ac.uk/wp-content/uploads/2022/02/220210-CVD-COVID-UK-COVID-IMPACT-TRE-Dataset-Provisioning-Dashboard.pdf) (last checked 11 February 2022).
  20. Who is at high risk from coronavirus (COVID-19). See [www.nhs.uk/conditions/coronavirus-covid-19/people-at-higher-risk/who-is-at-high-risk-from-coronavirus/](http://www.nhs.uk/conditions/coronavirus-covid-19/people-at-higher-risk/who-is-at-high-risk-from-coronavirus/) (last checked 1 February 2022).
  21. Denaxas S, Gonzalez-Izquierdo A, Direk K, Fitzpatrick NK, Fatemifar G, Banerjee A, et al. UK phenomics platform for developing and validating electronic health record phenotypes: CALIBER. *J Am Med Inform Assoc* 2019; 26: 1545–1559.
  22. OpenPrescribing, <https://openprescribing.net/bnf/0603/> (last checked 1 February 2022).
  23. Thygesen JH, Tomlinson C, Hollings S, Mizani MA, Handy A, Akbari A, et al. COVID-19 trajectories among 57 million adults in England: a cohort study using electronic health records. *Lancet Digit Health* 2022; 4: e542–e557.
  24. Estimates of the population for the UK, England and Wales, Scotland and Northern Ireland, Office for National Statistics. See [www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/populationestimatesforukenglandandwalesscotlandandnorthernireland](http://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/populationestimatesforukenglandandwalesscotlandandnorthernireland) (last checked 1 February 2022).
  25. Kopec JA, Finès P, Manuel DG, Buckeridge DL, Flanagan WM, Oderkirk J, et al. Validation of population-based disease simulation models: a review of concepts and methods. *BMC Public Health* 2010; 10: 710.
  26. Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *BMJ* 2020; 369: m1328.
  27. Clark A, Jit M, Warren-Gash C, Guthrie B, Wang HHX, Mercer SW, et al. Global, regional, and national estimates of the population at increased risk of severe COVID-19 due to underlying health conditions in 2020: a modelling study. *Lancet Glob Health* 2020; 8: e1003–e1017.
  28. Banerjee A, Pasea L, Manohar S, Lai AG, Hemingway E, Sofer I, et al. ‘What is the risk to me from COVID-19?’: Public involvement in providing mortality risk information for people with ‘high-risk’ conditions for COVID-19 (OurRisk.CoV). *Clin Med (Lond)* 2021; 21: e620–e628.
  29. Coronavirus (COVID-19) Infection survey: characteristics of people testing positive for COVID-19 in England and antibody data for the UK: December 2020. See [www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/articles/coronaviruscovid19infectionsinthecommunityinengland/december2020](http://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/articles/coronaviruscovid19infectionsinthecommunityinengland/december2020) (last checked 2 February 2022).
  30. Strongman H, Carreira H, De Stavola BL, Bhaskaran K and Leon DA. Factors associated with excess all-cause mortality in the first wave of the COVID-19 pandemic in the UK: a time series analysis using the Clinical Practice Research Datalink. *PLoS Med* 2022; 19: e1003870.